Stein's Method

Matthew Kwan

October 18, 2013

Contents

1 Introduction

Comment 1.1. The plan is to introduce with limit theorems: Central Limit theorem, Poisson Limit theorem. The failure of limit theorems is that they provide no understanding of speed of convergence, in particular convergence cannot be assumed to be uniform as parameters vary.

Stein's method is a technique for bounding the distance between distributions, with a variety of different distance metrics. Quantitative bounds can be useful in their own right, or can be further applied to prove asymptotic results.

1.1 Notation

For this thesis, the set of natural numbers $\mathbb{N}$ includes zero. We write 1_A for the characteristic function of A: $1_A(x) = 1$ if $x \in A$, otherwise $1_A(x) = 0$. Also, $[k]$ denotes the set $\{1, \ldots, k\}$.

Unless otherwise specified, all asympotics are as $n \to \infty$. Apart from standard asymptotic notation, we use two notions of asymptotic equivalence: $f \sim g$ means $f = g(1 + o(1))$ and $f \asymp g$ means $f = O(g)$ and $g = O(f)$.

Part I

Theory

2 General Probability Theory

2.1 Review of basic concepts

Comment 2.1. I'm a little bit uncertain how much depth to go into for this. At the moment, it's written so that someone who's seen measure theory but no probability theory (an analyst)

can understand. Where possible, I've tried to translate things into the discrete case, because it's often more intuitive (and since I plan for applications to be combinatorial).

For many combinatorial applications, an informal understanding of probability theory will suffice. However, in this thesis a rigorous foundation in probability theory will be useful. The following is intended only as a brief review.

Definition 2.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity. An *event* is a measurable set $A \in \mathcal{A}$.

Definition 2.2. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity. An *event* is a measurable set $A \in \mathcal{A}$.

For our purposes Ω will often be a finite set of combinatorial objects, with $\mathcal{A}$ as the power set of Ω. In this case $\mathbb{P}$ is defined by $\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, for each $\omega \in \Omega$. We will discuss specific probability spaces on combinatorial objects in Section 3, but we include a particularly useful definition here:

Definition 2.3. In a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where Ω is finite, if $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, then we say $\mathbb{P}$ is *uniform*.

For an event A, $\mathbb{P}(A)$ is interpreted as the "probability that A occurs". For combinatorial spaces, events are usually of the form $A = \{\omega \in \Omega : P(\omega) \text{ holds}\}$, where $P(\omega)$ is some property of an object ω. For clarity, we often abuse notation slightly and write $\mathbb{P}(P(\omega) \text{ holds})$ instead of $\mathbb{P}(A)$.

Definition 2.4. A *random element* $X : \Omega_1 \to \Omega_2$ is a measurable function from a probability space $(\Omega_1, \mathcal{A}_1, \mathbb{P})$ to some measure space $(\Omega_2, \mathcal{A}_2, \mu)$. If the target measure space is $\mathbb{R}^n$ with the Borel σ-algebra and the Lebesgue measure, then we say X is a *random vector*; if $n = 1$ then X is a *random variable*. If X only takes countably many values then we say X is *discrete*.

If the underlying probability space Ω_1 is countable, then any function is measurable.

We will often be interested in the probability that a random element takes certain values, without regard to the underlying probability space.

Definition 2.5. Suppose X is a random element with target measure space $(\Omega, \mathcal{A}, \mu)$. The *distribution* (or *law*) $\mathcal{L}_X$ of X is the pushforward measure with respect to X. That is, it is a probability measure defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for $A \subseteq \mathcal{A}$. We also occasionally use the notation $\mathcal{L}(X) := \mathcal{L}_X$ for ease of reading.

It is worth noting that in fact any probability measure is the distribution of some random element. To see this, note that given a probability measure $\mathbb{P} \in \mathcal{P}(\Omega)$, we can choose $X = \mathrm{id}_\Omega$ to have $\mathcal{L}_X = \mathbb{P}$. So, it is often convenient to specify random variables by their distributions, without defining an underlying probability space. We can use slightly abusive (but standard) notation like $\mathbb{P}(X > 1)$ to denote $\mathcal{L}_X(\{x : x > 1\})$. This is equal to $\mathbb{P}(\{\omega \in \Omega : X(\omega) > 1\})$ for any particular realization of X as a function on a probability space $(\Omega, \mathcal{A}, \mu)$.

Example 2.6. If X has the normal distribution with parameters μ and σ then we say $\mathcal{L}_X = \mathcal{N}(\mu, \sigma)$; this distribution is defined by $\mathcal{L}_X(B) = \frac{1}{\sigma\sqrt{2\pi}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}\, x$ for Borel B.

If X is discrete, then $\mathcal{L}_X$ is just an assignment of a probability to each possible value.

Example 2.7. If X is Poisson distributed with parameter λ, we write $\mathcal{L}_X = \mathrm{Po}(\lambda) = \mathrm{Po}_\lambda$; this is defined by $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

Definition 2.8. The *expected value* of a random variable X is $\mathbb{E}X = \int x \,\mathrm{d}\, \mathcal{L}_X(x)$.

For a random variable X that takes integer values, this definition is equivalent to the well-known formula $\mathbb{E}X = \sum_{x \in \mathbb{Z}} x\mathbb{P}(X = x)$.

If we fix a particular underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can also equivalently view expectation as a linear functional on the space of integrable functions: $\mathbb{E}X = \int X(\omega) \,\mathrm{d}\, \mathbb{P}$. So, $\mathbb{E}$ is defined in terms of a particular underlying probability space. Sometimes we will define a new probability space $(\Omega, \mathcal{A}, \mathbb{P}')$ by changing the measure on the same underlying set. In

this case we will write $\mathbb{E}_{\mathbb{P}'}$ to indicate expecation with respect to the measure $\mathbb{P}'$, to avoid ambiguity.

In fact, the expectation functional defines its underlying probability measure, because $\mathbb{E}\,1_A = \mathbb{P}(A)$. Since the distribution of a random variable is specified by a probability measure, the distribution $\mathcal{L}(X)$ of a random variable X also uniquely defines an expectation functional $\mathbb{E}_X := \mathbb{E}_{\mathcal{L}(X)}$.

Definition 2.9. For two collections $S, S' \subseteq \mathcal{A}_1$ of events, we say that S and S' are *independent* if $\mathbb{P}(A \cap A') = \mathbb{P}(A)\mathbb{P}(A')$ for each $A \in S$ and $A \in S'$. If $S = \{A\}$ contained a single set, then we say A itself is independent of S'.

Definition 2.10. Let $(\Omega_1, \mathcal{A}_1, \mathbb{P})$ be a probability space and $(\Omega_2, \mathcal{A}_2, \mu)$ a measure space. Let X be a random variable $\Omega_1 \to \Omega_2$, and let S be the set of all events of the form $\{\omega \in \Omega_1 : X(\omega) \in A_2\}$ for $A_2 \in \mathcal{A}_2$. If S is independent of S' then we say X itself is independent of S'.

We can analogously say that two random variables are independent, or a random variable and an event are independent, or any similar combination.

Definition 2.11. If two objects are not independent, then we say they are *dependent*.

Definition 2.12. Suppose $X : \Omega_1 \to \Omega_2$ is a random element defined on these spaces, and $A_1 \in \mathcal{A}_1$ is an event with nonzero probability. Then the *distribution of X conditioned on A_1* is denoted by $\mathcal{L}_{X|A_1}$ and defined by $\mathcal{L}_{X|A_1}(A_2) = \mathbb{P}(X \in A_2 | A_1)$ for $A_2 \in \mathcal{A}_2$. The expected value of a random variable with distribution $\mathcal{L}_{X|A_1}$ is called the *conditional expected value of X given A_1* and is denoted $\mathbb{E}[X|A_1]$.

We can also define conditional expectation with respect to another random variable. If X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then the sets $X_2^{-1}(B)$ for Borel B comprise a sub-σ-algebra $\mathcal{A}'$ of $\mathcal{A}$. Then, $\mu : A' \mapsto \mathbb{E}[X_1\, 1_{A'}]$ is a signed measure on $\mathcal{A}'$ that is absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}'$. By the Radon-Nikodym theorem there is an $\mathcal{A}'$-measurable random variable $\mathbb{E}[X_1|X_2]$ that satisfies $\mathbb{E}[X_1\, 1_{A'}] = \mathbb{E}[\mathbb{E}[X_1|X_2]\, 1_{A'}]$ for all A' in $\mathcal{A}'$. This random variable is almost uniquely defined: for any two choices of $\mathbb{E}[X_1|X_2]$, the probability that they differ is zero.

Definition 2.13. The random variable $\mathbb{E}[X_1|X_2]$ as defined above is called the *conditional expectation of X_1 with respect to X_2*. We can also view conditional expectation as a linear operator between functions: we define $\mathbb{E}^{X_2}$ by $X_1 \mapsto \mathbb{E}[X_1|X_2]$.

This definition generalizes the previous definition of expectation conditioned on an event: if $\omega \in A$ and $\mathbb{P}(A) > 0$ then $\mathbb{E}[X|\,1_A](\omega) = \mathbb{E}[X|A]$.

Note that if X_2 is discrete then we do not need to invoke Radon-Nikodym. We can define $\mathbb{E}[X_1|X_2]$ by $\mathbb{E}[X_1|X_2](\omega) = \mathbb{E}[X_1|X_2 = X_2(\omega)]$ whenever $\mathbb{P}(X_2 = X_2(\omega)) > 0$.

We finally present a simple consequence of the definition of conditional expectation.

Proposition 2.14 (Tower Law of Expectation)**.** $\mathbb{E}\big[\mathbb{E}^{X_2}X_1\big] = \mathbb{E}[X_1]$

Proof. $\mathbb{E}\big[\mathbb{E}^{X_2}X_1\big] = \mathbb{E}[\mathbb{E}[X_1|X_2]\,1_\Omega] = \mathbb{E}[X_1\,1_\Omega] = \mathbb{E}[X_1]$ □

2.2 Coupling

Given a finite collection of measure spaces $(\Omega_1, \mathcal{A}_1, \mu_1), \ldots, (\Omega_n, \mathcal{A}_n, \mu_n)$ recall the construction of the product measure space $(\Omega, \mathcal{A}, \mu) := (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \prod_{i=1}^n \mu_i)$. If a random element takes values in a product space then each component is measurable, and conversely if the components of a random tuple are measurable then that tuple is measurable in the product space. So, we can make the following definitions:

Definition 2.15. Given random elements $X_1, \ldots, X_n$ on the same underlying probability space, $\mathcal{L}(X_1, \ldots, X_n) := \mathcal{L}((X_1, \ldots, X_n))$ is called the *joint distribution* of $X_1, \ldots, X_n$. Conversely, given a random tuple $(X_1, \ldots, X_n)$, each $\mathcal{L}(X_i)$ is called a *marginal distribution*.

Suppose we have two distributions of random elements $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$. *Coupling* is the technique of constructing a random ordered pair (X_1, X_2) which realizes the given distributions as marginal distributions. Usually this is done by specifying the joint distribution $\mathcal{L}(X_1, X_2)$.

The idea is that coupling creates a particular kind of dependence between X_1 and X_2 that allows us to compare the two distributions. Often, we are able to make conclusions about the

distributions $\mathcal{L}(X_i)$ which are independent of their specific realizations as random elements in the coupling.

2.3 Markov Chains

Comment 2.2. I'll need to define Markov Chains, stationary distributions, irreducibility and time-reversibility.

Perhaps I should talk more generally about stochastic processes, because applying exchangeable pairs to Stein's method is has connections with Ornstein-Uhlenbeck processes and also Stein's method can be applied to Poisson processes.

2.4 The Weak Topology on Probability Measures

Comment 2.3. The main purpose of this section is to motivate the metrics usually used in Stein's method: they are all legitimate topological metrics and are consistent with the topology of convergence in distribution. In particular, if we can show $\mathrm{d}_{\mathcal{H}}(X_n, X) \to 0$ we have shown that $X_n \xrightarrow{d} X$, as Toby does [Joh11].

Definition 2.16. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say X_n *converges in distribution* to a random variable X if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f. Alternatively, we say $\mathcal{L}(X_n)$ converges *weakly* to $\mathcal{L}(X)$, or simply $\mathcal{L}(X_n) \to \mathcal{L}(X)$. The topology on $\mathcal{P}(\mathbb{R})$ associated with this convergence is called the *weak topology* (we will see that it is indeed a topology). Convergence in distribution of random vectors is defined componentwise.

Definition 2.17. The *distribution function* F_X of a random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$.

Theorem 2.18. *The following are equivalent.*

1. $\mathcal{L}(X_n) \to \mathcal{L}(X)$

2. $F_{X_n}(x) \to F_X(x)$ *for all* x *where* F_X *is continuous*

3. *(Lévy's continuity theorem)* $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ *for all* $t \in \mathbb{R}$.

The equivalence of Items 1 and 2 is a well-known result called the Portmanteau Theorem.

When X and each X_n are integer random variables, then Item 2 reduces to the condition that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. This characterization is usually used to prove the Poisson limit theorem. Lévy's continuity theorem is classically used to prove the central limit theorem, but we will not discuss it in this thesis.

For combinatorial applications, convergence in distribution can also be proved by the "method of moments": if X is the only random variable with the moments $\left(\mathbb{E}X^k\right)_{k\in\mathbb{N}}$, then $\mathcal{L}(X_n) \to \mathcal{L}(X)$ if $\mathbb{E}X_n^k \to \mathbb{E}X^k$. Convergence in distribution can also sometimes be inferred from stronger forms of convergence when X and all the X_n are coupled to the same underlying space.

A disadvantage of all these approaches is that it is difficult to quantify the rate of convergence.

In functional analysis terms, note that expectation operators are bounded linear functionals on the space of real bounded continuous functions. Then, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ just means that $\mathbb{E}_{X_n} \to \mathbb{E}_X$ in the weak-star topology. Although $C_b(\mathbb{R})^*$ is not metrizable, the subspace corresponding to $\mathcal{P}(\mathbb{R})$ is in fact metrizable, with a metric called the Lévy metric. For Stein's method we will be interested in some slightly stronger metrics.

Definition 2.19. For two probability measures $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R})$ and a collection of real measurable "test" functions $\mathcal{H}$, define $d_{\mathcal{H}}$ by $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{\mathbb{P}_1} h - \mathbb{E}_{\mathbb{P}_2} h|$. For random variables X_1, X_2, we write $d_{\mathcal{H}}(X_1, X_2)$ instead of $d_{\mathcal{H}}(\mathcal{L}(X_1), \mathcal{L}(X_2))$.

Each $d_{\mathcal{H}}$ is non-negative, symmetric and satisfies the triangle inequality.

Definition 2.20. A set of real functions $\mathcal{H}$ is a *determining class* if $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$ implies that $\mathbb{P}_1 = \mathbb{P}_2$.

To check that $d_{\mathcal{H}}$ is a metric, we only need to check that $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ implies that $\mathbb{P}_1 = \mathbb{P}_2$. That is, we need to check that $\mathcal{H}$ is a determining class.

Definition 2.21. We define some special cases of $d_{\mathcal{H}}$.

- If $\mathcal{H}_{\mathrm{K}} = \left\{1_{(-\infty,x]} : x \in \mathbb{R}\right\}$ then $d_{\mathrm{K}} := d_{\mathcal{H}_{\mathrm{K}}}$ is called the *Kolmogorov metric.*
- If $\mathcal{H}_{\mathrm{W}}$ is the set of real functions h that satisfy $|h(x_1) - h(x_2)| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$ (that is, the set of functions with Lipschitz constant 1), then $d_{\mathrm{W}} := d_{\mathcal{H}_{\mathrm{W}}}$ is called the *Wasserstein metric.*
- If $\mathcal{H}_{\mathrm{TV}}$ is the set of functions 1_B for Borel B, $d_{\mathrm{TV}} := d_{\mathcal{H}_{\mathrm{TV}}}$ is called the *total variation metric.*

Proposition 2.22. *The Kolmogorov, Wasserstein and total variation "metrics" are actually metrics.*

Proof. We check that $\mathcal{H}_{\mathrm{K}}$, $\mathcal{H}_{\mathrm{W}}$ and $\mathcal{H}_{\mathrm{TV}}$ are determining classes. Let $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{W}}, \mathcal{H}_{\mathrm{TV}}\}$, and suppose that $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$. It suffices to prove that $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$, since the sets $(-\infty, x]$ generate the Borel σ-algebra. For $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{TV}}\}$ this is immediate, because $\mathbb{P}_i((-\infty, x]) = \mathbb{E}_{\mathbb{P}_i} 1_{(-\infty,x]}$. So, consider, $\mathcal{H} = \mathcal{H}_{\mathrm{W}}$.

For $\varepsilon > 0$ and $x \in \mathbb{R}$, let $h_{x,\varepsilon}$ be the continuous function which takes the value 1 on the set $(-\infty, x]$, takes the value 0 on the set $[x + \varepsilon, \infty)$, and is linearly interpolated in the range $[x, x + \varepsilon]$. Since $\varepsilon h_{x,\varepsilon} \in \mathcal{H}_{\mathrm{W}}$, we have $\mathbb{E}_{\mathbb{P}_1} h_{x,\varepsilon} = \mathbb{E}_{\mathbb{P}_2} h_{x,\varepsilon}$ for each $n \in \mathbb{N}$. For each $x \in \mathbb{R}$, $h_{x,1/n} \to 1_{(-\infty,x]}$ pointwise and each $h_{x,1/n} \leq 1$ so by the dominated convergence theorem, $\mathbb{E}_{\mathbb{P}_i} h_{1/n} \to \mathbb{E}_{\mathbb{P}_i} 1_{(-\infty,x]}$ for each $i \in \{1, 2\}$. We have again proved that $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$. □

Proposition 2.23. *The topologies induced by the Kolmogorov, Wasserstein and total variation metrics are each stronger than the weak topology.*

Proof. If $d_{\mathrm{K}}(X_n, X) \to 0$ or $d_{\mathrm{TV}}(X_n, X) \to 0$ then $F_{X_n} \to F_X$ uniformly, so certainly Item 2 of Theorem 2.18 holds.

Now, suppose $d_{\mathrm{K}}(X_n, X) \to 0$. Let $d_n = \sqrt{d_{\mathrm{K}}(X_n, X)}$ and recall the definition of $h_{x,\varepsilon}$ from the proof of Proposition 2.22. Since $d_n h_{x,d_n} \in \mathcal{H}_{\mathrm{K}}$ for each $n \in \mathbb{N}$, we have $\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n} \leq d_{\mathrm{K}}(X_n, X)/d_n = d_n \to 0$ uniformly for $x \in \mathbb{R}$. Now, note that $F_X(x-\varepsilon) \leq \mathbb{E}_X h_{x-\varepsilon,\varepsilon} \leq F_X(x) \leq \mathbb{E}_X h_{x,\varepsilon} \leq F_X(x+\varepsilon)$ for any random variable X. If F_X is continuous at x then

$$F_{X_n}(x) - F_X(x) \leq (\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n}) + (F_X(x+d_n) - F_X(x)) \to 0$$
$$F_{X_n}(x) - F_X(x) \geq (\mathbb{E}_{X_n} h_{x-d_n,d_n} - \mathbb{E}_X h_{x-d_n,d_n}) + (F_X(x-d_n) - F_X(x)) \to 0$$

so Item 2 of Theorem 2.18 holds. □

Proposition 2.23 tells us that we can sensibly use our metrics to quantify the distance between random variables, in a way that is consistent with distributional (weak) convergence. All three metrics are relevant in their own right, but sometimes one may be easier to work with. It is sometimes possible to transfer results between metrics, though this usually results in worse constants than working directly in the desired metric.

Issue 2.4. It may be worthwhile to actually characterize the Wasserstein, Kolmogorov and Total Variation topologies. In particular, Wikipedia says that Wasserstein convergence is just weak convergence plus convergence of the first moment.

Definition 2.24. If $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}\,x$ then f_X is called the *Lebesgue density* of X, and X is called a *continuous* random variable.

If X is a continuous random variable, then by the Radon-Nikodym chain rule $\mathbb{E}_X h = \int_{\mathbb{R}} h(x) f_X(x)\,\mathrm{d}\,x$.

Proposition 2.25. *Let X_1, X_2 be random variables.*

1. $d_{\mathrm{K}}(X_1, X_2) \leq d_{\mathrm{TV}}(X_1, X_2)$
2. *If X_2 has Lebesgue density bounded by C, then $d_{\mathrm{K}}(X_1, X_2) \leq \sqrt{2C d_{\mathrm{W}}(X_1, X_2)}$.*

Proof. (Adapted from [Ros11, Proposition 1.2]). Item 1 is immediate from the definition. Then, as in the proof of Proposition 2.23,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,\varepsilon} - \mathbb{E}_X h_{x,\varepsilon}) + (\mathbb{E}_X h_{x,\varepsilon} - F_X(x)) \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + \int_x^{x+\varepsilon} h_{x,\varepsilon} f_X(x) \,\mathrm{d}\, x \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + C\varepsilon/2 \end{aligned}$$

and similarly

$$F_{X_n}(x) - F_X(x) \geq -d_{\mathrm{W}}(X_1, X_2)/\varepsilon - C\varepsilon/2,$$

So, we can take $\varepsilon = \sqrt{2d_{\mathrm{W}}(X_1, X_2)/C}$ to prove Item 2. □

Example 2.26. If $\mathcal{L}_{X_2} = \mathcal{N}(0,1)$ then $d_{\mathrm{K}} \leq (2/\pi)^{1/4}\sqrt{d_{\mathrm{W}}(X_1, X_2)}$.

In a combinatorial setting, many of our results are about integer random variables. The total variation metric is usually exclusively used in this case.

Proposition 2.27. *If X_1, X_2 are integer-valued random variables, then*

$$d_{\mathrm{TV}}(X_1, X_2) = \frac{1}{2}\sum_{k\in\mathbb{Z}} |\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)|.$$

Proof. For any Borel set A, let $d_A = \mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)$, so that $d_{\mathrm{TV}}(X_1, X_2) = \sup|d_A|$. Define

$$\begin{aligned} A_< &= \{k \in \mathbb{Z} : \mathbb{P}(X_1 = k) < \mathbb{P}(X_2 = k)\}, \\ A_> &= \{k \in \mathbb{Z} : \mathbb{P}(X_1 = k) > \mathbb{P}(X_2 = k)\}. \end{aligned}$$

For any Borel A, we have

$$\begin{aligned} d_A &= \sum_{k \in \mathbb{Z} \cap A} (\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)) \\ &\leq \sum_{k \in A_>} (\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)) \\ &= d_{A_>} \end{aligned}$$

and similarly $\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A) \geq d_{A_<}$. Since $d_{A_>} = -d_{A_<}$, we have

$$d_{\mathrm{TV}}(X_1, X_2) = (d_{A_>} - d_{A_<})/2 = \frac{1}{2} \sum_{k \in \mathbb{Z}} |\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)|.$$

□

3 Random Combinatorial Structures

Comment 3.1. I'll leave this section until I'm sure what applications I'll be looking at. Definitely random graph models, probably random permutations and random matrices.

4 Stein's Method in the Abstract

Comment 4.1. There are a few quite different presentations of Stein's method. One thing I'm trying to do here is to unify Stein's functional analysis approach for exchangeable pairs [Ste86] with Ross' general presentation[Ros11].

The reason I want to look at Stein's original, more abstract presentation is that I think it does a better job motivating why things work. Before I read that, the steps taken to apply Stein's method seemed like blindly doing things and it turns out they work.

Suppose we have a potentially complicated random variable X, and we believe the distribution of X is close to a "standard" distribution $\mathcal{L}_0$. Then, Stein's method allows us to compare the operators $\mathbb{E}_X$ and $\mathbb{E}_0 := \mathbb{E}_{\mathcal{L}_0}$. This is sometimes directly useful for approximating statistics of X (for example, $\mathbb{P}(X \in A) = \mathbb{E}_X 1_A$). However, particularly for combinatorical applications, Stein's method is most often used to bound the distance $d_{\mathcal{H}}(\mathcal{L}_X, \mathcal{L}_0)$, where the metric $d_{\mathcal{H}}$ from Definition 2.19 is defined in terms of $\mathbb{E}_X$ and $\mathbb{E}_0$.

Stein's method is motivated by the idea of a characterizing operator.

Definition 4.1. Let $\mathcal{F}_0$ be a vector space and $\mathcal{X}_0$ be a vector space of measurable functions. We say a linear operator $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a *characterizing operator* for the distribution $\mathcal{L}_0$ if $\operatorname{im} T_0 = \mathcal{X}_0 \cap \ker \mathbb{E}_0$. For convenience, where there is no ambiguity we will often implicitly restrict $\mathbb{E}_0$ to $\mathcal{X}_0$, so we can write $\operatorname{im} T_0 = \ker \mathbb{E}_0$.

The following proposition shows why T_0 is called a characterizing operator.

Proposition 4.2. *If $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a characterizing operator and $\mathcal{X}_0$ is a determining class then $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_X$ implies $\mathcal{L}_X = \mathcal{L}_0$.*

Proof. If $h \in \mathcal{X}_0$, then $h - \mathbb{E}_0 h \in \ker \mathbb{E}_0 = \operatorname{im} T_0$ so $\mathbb{E}_X[h - \mathbb{E}_0 h] = 0$. That is, $\mathbb{E}_X h = \mathbb{E}_0 h$ for all $h \in \mathcal{X}_0$, which means $\mathcal{L}_X = \mathcal{L}_0$ by the definition of a determining class. □

Issue 4.2. Ross [Ros11] and others use this weaker condition as the definition of a characterizing operator. I'll have to look at examples of operators that satisfy the weaker but not the stronger condition to see if the stronger definition is warranted (my guess is yes, if Stein decided to originally define it the way I did).

Proposition 4.3. *$T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is characterizing if and only if there is a linear operator $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$ such that the following two equations hold.*

$$\mathbb{E}_0 T_0 = 0_{\mathcal{F}_0}, \tag{4.1}$$

$$T_0 U_0 + \mathbb{E}_0 = \mathrm{id}_{\mathcal{X}_0}. \tag{4.2}$$

Proof. Suppose T_0 is a characterizing operator. Equation (4.1) is immediate. Let $\{h_i\}_{i\in\mathcal{I}}$ be a (Hamel) basis of $\mathcal{X}_0$. For each $i \in \mathcal{I}$ we have $h_i - \mathbb{E}_0 h_i \in \ker \mathbb{E}_0$ so there is some f_i (not necessarily unique) that solves $T_0 f_i = h_i - \mathbb{E}_0 h_i$. The operator U_0 can then be defined by $\sum_{i\in\mathcal{I}} a_i h_i \mapsto \sum_{i\in\mathcal{I}} a_i f_i$, satisfying (4.2).

Issue 4.3. there's probably a cleaner functional analysis way to prove that. Also, is U_0 bounded?

Conversely, suppose (4.1) holds and U_0 exists satisfying (4.2). For $h \in \ker \mathbb{E}_0$ we have $T_0(U_0 h) = h$ and $h \in \operatorname{im} T_0$, so $\ker \mathbb{E}_0 \subseteq \operatorname{im} T_0$. Equation (4.1) immediately says that $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_0$, so T_0 is a characterizing operator. □

We'll use Proposition 4.3 to give two important examples of characterizing operators.

Theorem 4.4. *Define $T_{\mathcal{N}}$ by $T_{\mathcal{N}} f(x) = f'(x) - xf(x)$. Let $\mathcal{X}_{\mathcal{N}}$ be the set of functions $h : \mathbb{R} \to \mathbb{R}$ that satisfy $\mathbb{E}_{\mathcal{N}}|h| < \infty$ and let $\mathcal{F}_{\mathcal{N}}$ be the set of functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{\mathcal{N}}|T_{\mathcal{N}} f| < \infty$. Then $T_{\mathcal{N}} : \mathcal{F}_{\mathcal{N}} \to \mathcal{X}_{\mathcal{N}}$ is a characterizing operator for $\mathcal{N}(0,1)$.*

Proof. For any $f \in \mathcal{F}_{\mathcal{N}}$, integration by parts gives

$$\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} f'(t) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2/2} f(t) \,\mathrm{d}\, t = 0$$

so $\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} = 0$ and (4.1) holds. Then, define $U_{\mathcal{N}}$ by

$$U_{\mathcal{N}} h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t.$$

By the product rule and the fundamental theorem of calculus, for all $h \in \mathcal{X}_{\mathcal{N}}$ we have

$$T_{\mathcal{N}} U_{\mathcal{N}} h(x) = h(x) - \mathbb{E}_{\mathcal{N}} h,$$

so (4.2) holds and Proposition 4.3 completes the proof. □

Theorem 4.5. *Define* $T_{\mathrm{Po}(\lambda)}$ *by* $T_{\mathrm{Po}(\lambda)}f(k) = \lambda f(k+1) - kf(k)$. *Let* $\mathcal{X}_{\mathrm{Po}(\lambda)}$ *be the set of integer-valued functions* $h : \mathbb{N} \to \mathbb{Z}$ *that satisfy* $\mathbb{E}_{\mathrm{Po}(\lambda)}|h| < \infty$ *and let* $\mathcal{F}_{\mathrm{Po}(\lambda)}$ *be the set of functions* $f : \mathbb{N} \to \mathbb{Z}$ *such that* $\mathbb{E}_{\mathrm{Po}(\lambda)}\left|T_{\mathrm{Po}(\lambda)}f\right| < \infty$. *Then* $T_{\mathrm{Po}(\lambda)} : \mathcal{F}_{\mathrm{Po}(\lambda)} \to \mathcal{X}_{\mathrm{Po}(\lambda)}$ *is a characterizing operator for* $\mathrm{Po}(\lambda)$.

Proof. For any $f \in \mathcal{F}_{\mathrm{Po}(\lambda)}$, we have

$$\mathbb{E}_{\mathrm{Po}(\lambda)}T_{\mathrm{Po}(\lambda)}f = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^{i+1}}{i!}f(i+1) - e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i}}{(i-1)!}f(i) = 0$$

so $\mathbb{E}_{\mathrm{Po}(\lambda)}T_{\mathrm{Po}(\lambda)} = 0$ and (4.1) holds. Then, define $U_{\mathrm{Po}(\lambda)}$ by

$$U_{\mathrm{Po}(\lambda)}h(k) = \frac{(k-1)!}{\lambda^k}\sum_{i=0}^{k-1}\frac{\lambda^i}{i!}\big(h(i) - \mathbb{E}_{\mathrm{Po}(\lambda)}h\big)$$

for $k \geq 1$. Substituting and simplifying gives

$$T_{\mathrm{Po}(\lambda)}U_{\mathrm{Po}(\lambda)}h(k) = h(k) - \mathbb{E}_{\mathrm{Po}(\lambda)}h,$$

so (4.2) holds and Proposition 4.3 completes the proof. □

Note that $\mathcal{H}_{\mathrm{TV}} \subseteq \mathcal{X}_{\mathcal{N}}$, where $\mathcal{H}_{\mathrm{TV}}$ is as defined in Definition 2.21. Since $\mathcal{H}_{\mathrm{TV}}$ is a determining class, $T_{\mathcal{N}}$ is a characterizing operator in the sense of Proposition 4.2. We can say the same about $T_{\mathrm{Po}(\lambda)}$ if we restrict our attention to integer-valued random variables.

Issue 4.4. Stein chose $\mathcal{X}_0 = \big\{h : \mathbb{E}\big[\mathrm{id}_{\mathbb{R}}^k|h|\big] < \infty \text{ for all } k\big\}$, for both the Poisson and normal case. I'm not sure why, I'll revisit this after looking at exchangeable pairs.

The utility of the introduction of a characterizing operator is that for each $h \in \mathcal{X}_0$, Equation (4.2) allows us to to make the transformation

$$\mathbb{E}_X h = \mathbb{E}_0 h + \mathbb{E}_X T_0 U_0 h. \tag{4.3}$$

The original purpose of Stein's method was to estimate some particular $\mathbb{E}_X h$. If $\mathcal{L}_0$ was chosen to be a "simple", well-understood distribution then the term $\mathbb{E}_0 h$ should be easy to compute or estimate, and if the distribution of X was "close" to $\mathcal{L}_0$, then it should be possible to show that the remainder $\mathbb{E}_X T_0 U_0 h$ is small.

For our purposes, the main use of (4.3) is to bound $d_{\mathcal{H}}(X, \mathcal{L}_0)$ for some $\mathcal{H} \subseteq \mathcal{X}_0$. If $\mathcal{Y} \subseteq \mathcal{F}_0$ is chosen so that $U_0 h \in \mathcal{Y}$ for all $h \in \mathcal{H}$, then we have

$$d_{\mathcal{H}}(X, \mathcal{L}_0) = \sup_{h \in \mathcal{H}} |\mathbb{E}_X T_0 U_0 h| \leq \sup_{f \in \mathcal{Y}} |\mathbb{E}_X T_0 f|.$$

We have reduced the problem of bounding $d_{\mathcal{H}}(X, \mathcal{L}_0)$ to that of bounding $|\mathbb{E}_X T_0 f|$ (uniformly over $f \in \mathcal{Y}$). Especially in the cases where $\mathcal{L}_0$ is normal or Poisson and $\mathcal{H}$ is one of the standard choices in Definition 2.21, there are a number of known convenient choices of $\mathcal{Y}$, and a number of methods that are known to be effective to bound $|\mathbb{E}_X T_0 f|$.

Example 4.6. If $\mathcal{H} = \mathcal{H}_{\mathrm{TV}}$ and $\mathcal{L}_0 = \mathrm{Po}(\lambda)$, then we can choose $\mathcal{Y} = \big\{ f \in \mathcal{F}_0 : \|f\|_\infty \leq 2 \min\big\{1, \lambda^{-1/2}\big\} \big\}$. This choice is sharp in the sense that $\sup_A \big\| U_{\mathrm{Po}(\lambda)} 1_A \big\|_\infty \asymp \lambda^{-1/2}$ as $\lambda \to \infty$ and $\sup_A \big\| U_{\mathrm{Po}(\lambda)} 1_A \big\|_\infty \asymp 1$ as $\lambda \to 0$.

Proof. (Adapted from [BHJ92, Lemma 1.1.1])... □

Issue 4.5. [BHJ92, Remark 10.2.4] gives a proof that $\|f\|_\infty \leq \min\big\{1, \lambda^{-1/2}\big\}$ will suffice, but it's complicated. Maybe I'll revisit this later.

Issue 4.6. I should have a toy example here that is amenable to several methods.

4.1 The method of Exchangeable Pairs

This is Stein's original approach, and is effective in wide generality for combinatorial random variables. In what follows, we assume $\Omega = \mathcal{A}$ is a finite set (presumably of combinatorial

objects) and $X : \Omega \to \mathbb{R}$ is a random variable that represents some statistic of the objects in Ω. The set $\mathbb{R}_X = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$ is then finite. We write its elements as $x^{(1)} < \cdots < x^{(k)}$ for $k = |\mathbb{R}_X|$.

The general idea is that we can use an object $\mathbf{X}$ called an exchangeable pair to construct a characterizing operator $T_{\mathbf{X}}$ for X. We then use an operator α to connect the domains of $T_{\mathbf{X}}$ and T_0 in such a way that $T_{\mathbf{X}}\alpha$ approximates T_0. We then have $\mathbb{E}_X T_0 = \mathbb{E}_X(T_0 - T_{\mathbf{X}}\alpha)$, so we can use the fact that $T_0 - T_{\mathbf{X}}\alpha$ is small to bound $\mathbb{E}_X T_0 f$.

Issue 4.7. This finiteness condition is ugly. Stein's book [Ste86] says "presumably an analogous result holds even if Ω is infinite", but his later paper [Ste92] makes no mention of this. Here's what needs to be adjusted to let Ω be infinite:

1. If Ω is uncountable, the definition $T_X f(x) = \mathbb{E}[f(X, X')|X = x]$ is no longer appropriate. There are ugly fixes, possibly something neater can be done with Radon-Nikodym. Stuff can be reformulated to define $T_X = \mathbb{E}^X$ but this would obscure the motivation of T_X as an approximation to T_0.
2. $\mathcal{F}$ and $\mathcal{X}$ need to be adjusted. This is probably not a big deal.
3. The proof that $\operatorname{im} T_X = \ker \mathbb{E}^X$ would no longer work. Apparently the result is related to homology, so I might look into that. The problem might go away if I use the alternative definition of a characterizing operator as in Issue 4.2.

Definition 4.7. A 2-dimensional random vector $\mathbf{X} = (X_1, X_2)$ is an *exchangeable pair* if $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$.

That is, a pair $\mathbf{X}$ is exchangeable if exchanging the components of the pair does not change their joint distribution. In particular, the marginal distribution of X_1 and X_2 must be the same. We will be interested in the case when $\mathcal{L}_{X_1} = \mathcal{L}_{X_2} = \mathcal{L}_X$.

Comment 4.8. All presentations of Stein's method I've seen use the notation (X, X') but I think that has the potential to be confusing because the X in that pair is defined on Ω^2 whereas the original random variable X is defined on Ω.

Proposition 4.8. *There is a natural equivalence between time-homogeneous reversible Markov chains with steady-state distribution $\mathcal{L}_X$, and exchangeable pairs with margins $\mathcal{L}_X$.*

Proof. Given an exchangeable pair $\mathbf{X}$ with margins $\mathcal{L}_X$, we can define a time-homogenous Markov chain M with transition probabilities $p_{ij} = \mathbb{P}\big(X_2 = x^{(j)} | X_1 = x^{(i)}\big)$. With $\pi_i = \mathbb{P}\big(X_1 = x^{(i)}\big)$, we then have

$$\pi_i p_{ij} = \mathbb{P}(\mathbf{X} = (x_i, x_j)) = \pi_j p_{ji},$$

so M is reversible with steady-state distribution $\mathcal{L}_X$.

Conversely, suppose a time-homogeneous Markov chain with transition probabilities p_{ij} satisfies $\pi_i p_{ij} = \pi_j p_{ji}$ for $\pi_i = \mathbb{P}\big(X_1 = x^{(i)}\big)$. We can then define an exchangeable pair $\mathbf{X}$ by

$$\mathbb{P}\Big(\mathbf{X} = \Big(x^{(i)}, x^{(j)}\Big)\Big) = \pi_i p_{ij} = \pi_j p_{ji} = \mathbb{P}\Big(\mathbf{X} = \Big(x^{(j)}, x^{(i)}\Big)\Big).$$

□

For any exchangeable pair $\mathbf{X}$, we can also define its graph $G_{\mathbf{X}}$ in a natural way: the vertex set $[k]$ corresponds to $\mathbb{R}_X = \big\{x^{(1)}, \ldots, x^{(k)}\big\}$, and there is an edge for each possible transition in the Markov chain of $\mathbf{X}$. We will be particularly interested in the interpretation of $G_{\mathbf{X}}$ as an (abstract) simplicial complex. That is, the set of one-dimensional faces F_1 is the edge set of $G_{\mathbf{X}}$ and the set of zero-dimensional faces F_0 is the vertex set $[k]$ of $G_{\mathbf{X}}$.

Definition 4.9. We say an exchangeable pair $\mathbf{X}$ is *connected* if its Markov chain is irreducible, or equivalently $G_{\mathbf{X}}$ is connected.

Especially when $\mathbb{P}$ is a uniform probability on a set of combinatorial objects, it is often useful to construct an exchangeable pair from a reversible Markov chain on Ω. Specifically, if we choose a reversible Markov chain with steady-state distribution given by $\pi = \mathbb{P}(\omega_i)$, then the vector $(X(W_1), X(W_2))$ is an exchangeable pair with margins $\mathcal{L}_X$. Further, if the underlying Markov chain is irreducible, then the constructed exchangeable pair is connected. We will frequently make use of this construction in applications.

Now, if $\mathbf{X}$ is a connected exchangeable pair with margins $\mathcal{L}_X$, consider the augmented chain complex of $G_{\mathbf{X}}$ over $\mathbb{R}$:

$$0 \to \mathbb{R}^{F_1} \xrightarrow{\partial_1} \mathbb{R}^{F_0} \xrightarrow{\partial_0} \mathbb{R} \to 0.$$

Since $G_{\mathbf{X}}$ is connected, its zeroth homology group has dimension zero. That is, $\operatorname{im}\partial_1 = \ker\partial_0$. This motivates the construction of a characterizing operator for X.

Let $\mathcal{F}_X$ be the set of functions $\mathbb{R}_X^2 \to \mathbb{R}$ that are antisymmetric in the sense that $f(x_1, x_2) = -f(x_2, x_1)$, and let $\mathcal{X}_X$ be the set of functions $\mathbb{R}_X \to \mathbb{R}$.

Theorem 4.10. *Let $\mathcal{F}_X$ and $\mathcal{X}_X$ be defined as above. Suppose $\mathbf{X}$ is a connected exchangeable pair with margins $\mathcal{L}_X$. Define $T_{\mathbf{X}} : \mathcal{F}_X \to \mathcal{X}_X$ by $T_{\mathbf{X}} f(x) = \mathbb{E}[f(X_1, X_2)|X_1 = x]$, so that $T_{\mathbf{X}} X = \mathbb{E}^{X_1} f(X_1, X_2)$. Then $T_{\mathbf{X}}$ is a characterizing operator for X.*

Proof. To see that $\operatorname{im} T_{\mathbf{X}} \subseteq \ker \mathbb{E}_X$, fix $f \in \mathcal{F}$ and note that by the tower law of expectation (Proposition 2.14),

$$\mathbb{E}_X T_{\mathbf{X}} f = \mathbb{E}\mathbb{E}^{X_1} f(X_1, X_2) = \mathbb{E} f(X_1, X_2).$$

By exchangeability and antisymmetry, $\mathbb{E} f(X_1, X_2) = \mathbb{E} f(X_2, X_1) = -\mathbb{E} f(X_1, X_2)$, so $\mathbb{E} f(X_1, X_2) = \mathbb{E}_X T_{\mathbf{X}} f = 0$. This did not require the connectedness condition.

We next prove $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$ using the fact that $\operatorname{im}\partial_1 = \ker\partial_0$. A chain in $\mathbb{R}^{F_i}$ will be represented as a function $F_i \to \mathbb{R}$.

Let $h \in \ker \mathbb{E}_X$. That is, $\sum_{i=1}^k h\big(x^{(i)}\big)\pi_i = 0$, so if we define $h^* : i \mapsto h\big(x^{(i)}\big)\pi_i$, then $h^* \in \ker\partial_0 = \operatorname{im}\partial_1$. This means there is $f^* \in \mathbb{R}^{F_1}$ with $\partial_1 f^* = h^*$, or equivalently

$$h^*(i) = \sum_{j \in [k] \setminus \{i\}} f^*(\{i, j\}) \operatorname{sign}(i - j)$$

for each i. Now, define $f : \mathbb{R}_X^2 \to \mathbb{R}$ by

$$f\Big(x^{(i)}, x^{(j)}\Big) = \begin{cases} \dfrac{f^*(\{i,j\})\operatorname{sign}(i-j)}{\pi_i p_{ij}} & \text{if } \{i,j\} \in F_1 \\ 0 & \text{otherwise.} \end{cases}$$

By the reversibility of the Markov chain of $\mathbf{X}$, the function f is antisymmetric so $f \in \mathcal{F}_X$. By construction, we have

$$h(i) = \sum_{j \in [k] \setminus \{i\}} f\Big(x^{(i)}, x^{(j)}\Big) p_{ij} = T_{\mathbf{X}} f(i)$$

for each i, so $h \in \operatorname{im} T_{\mathbf{X}}$. □

4.2 Size-Bias Coupling

Part II

Applications

Comment 4.9. I'd like to go into a number of small examples (perhaps interspersed in the discussion of Stein's method in Part I), but I'd like to also go through a number of "big" examples. I'd like these examples to showcase

- different types of results: most applications give quantitative estimates. [Joh11] gives a non-quantitative distributional convergence result that was not previously proved using other methods. There are also results that have no connection with distribution metrics, such as the concentration inequalities in [Ros11]. In particular, the Latin rectangle example in [Ste86] is interesting in that the final result is not probabilistic.
- different types of distributions: definitely at least the Poisson and normal case, perhaps also an example of a more exotic distribution like the one in [FG12] or perturbations of Poisson/normal distributions as in [BČX07].

- different ways to apply stein's method: definitely exchangeable pairs and probably size-biasing. Maybe also Zero-bias coupling.

References

[BČX07] Andrew D Barbour, Vydas Čekanavičius, and Aihua Xia, *On stein's method and perturbations*, arXiv preprint math/0702008 (2007).

[BHJ92] Andrew D Barbour, Lars Holst, and Svante Janson, *Poisson approximation*, Clarendon press Oxford, 1992.

[FG12] Jason Fulman and Larry Goldstein, *Stein's method and the rank distribution of random matrices over finite fields*, arXiv preprint arXiv:1211.0504 (2012).

[Joh11] Tobias Johnson, *Exchangeable pairs, switchings, and random regular graphs*, arXiv preprint arXiv:1112.0704 (2011).

[Ros11] Nathan Ross, *Fundamentals of stein's method*, Probab. Surv **8** (2011), 210–293.

[Ste86] Charles Stein, *Approximate computation of expectations*, Lecture Notes – Monograph Series **7** (1986).

[Ste92] ———, *A way of using auxiliary randomization*, Probability Theory (1992), 159–180.